Name: Graham Northrup
Section: 1

TTIC 31050 # 3

Sequences

The five sequences were are using to investigate BLAST parameters are

- Acinotebacter Phase AP205 (133 residues)
- Raptor Adenovirus head domain (137 residues)
- E. Coli CysK (323 residues)
- E. Coli EC536 toxin (239 residues)
- Danio STRA6 (670 residues)

Basic Parameter BLAST

The default parameters for NCBI BLAST are a word size of 6, and the BLOSUM62 scoring matrix. For the first and second sequences, we saw BLAST return several identical strains and then some with lower identity, as low as 30% for the phage protein, and 43% for the adenovirus. The returns for the two E. Coli sequences were many database entries of the same protein (as the database aggregates these). All of these alignments had E values of 0 (or arbitrarily close to it), and 99% identity. The fifth sequence returned all alignments with E score 0, but lower levels of identity (towards 50%). Each of these sequences were named similarly and were scoring very close to their max score.

Word Size

The NCBI BLAST tool offers three words sizes, the default 6, as well as 3 and 2.

Word Size = 3

For this word size, we saw more sequences returned for the BLAST against our first and second sequence. These new sequences had a wider variety of identities (as low as 24%) and had more variety (read larger) in E value. The E. Coli sequence results were largely unimpacted by this new word size. The Danio sequence returned potentially more sequences, some with lower identity scores. But in all three cases the E values appeared unaffected and all stayed very close to 0.

Word Size = 2

The effects for the word size of 2 were rather analogous to the word size of 3. However I did have much more trouble running the BLAST to completion without encountering an error during these tests (this was while using BLOSUM62). Perhaps using a different scoring matrix will yield different results for this word size to distinguish between a word size of 3 and 2.

Summary of effects

It appears that decreasing the word size "widens the net" on the BLAST search. We can see this as several of our searches returned more sequences (with by most measures 'worse' metrics than the sequences that appeared in the 6 and the shorter word length searches. Another way we can see this that doesn't necessarily reflect in the numbers is the runtimes of the BLASTs. With word length 6, the searches ran relatively quickly in each case (longer sequences generally taking longer). With shorter word lengths, the searches took longer (and in some cases crashed out and required rerunning the search, in particular with word size 2).

Scoring Matrix

The NCBI BLAST tool offers several PAM and BLOSUM matrices.

BLOSUM90

The main thing that I noticed when using BLOSUM90 for the scoring matrix is all of sequences returned in all 5 cases had extremely low E values. This is likely because BLOSUM90 is created using very homologous sequences and this is reflected in the results from the BLAST.

BLOSUM45

I next used the other extreme of the BLOSUM matrices, BLOSUM45. Compared the BLOSUM90 the data used to construct BLOSUM45 was much messier and there was a lot less similarity. This resulted in more varied sequences being returned by the BLAST. We see the return of larger E values in the first two queries, running all the way to 10 (the exact E value threshold set by NCBI BLAST). Unfortunately again it is hard to see any difference in the last three queries because of the number of exact matches that can be returned on each query

PAM30

For PAM matrices I plan to use both extremes offered by the tool, PAM30 and PAM250. Starting with PAM30 we see very similar results to BLOSUM90 as we can recall that lower numbers for PAM matrices mean more stringent conditions on the data used to seed the matrix. We lose all of the E values except for those exceptionally close to 0, in all 5 sequences.

PAM250

When using PAM250 we see the first and second query behave slightly differently. The first query still only returns ¿95% identity sequences with very low E values. However the adenovirus BLAST using PAM250 returns the same high similarity sequences as well as a group fo Turkey adenovirus results with 24% similarity and higher E values, all above 6. I expected both queries to return high E value sequences, as we saw with BLOSUM45, however I am not why we would only see this in one of the two queries here for PAM250.

Combined Effects

I now want to see how combining changes in these parameters will impact the results. I will first try and cast the net as wide as possible as use a lax matrix with a small word size.

Word Size=2, BLOSUM45

As expected this set of parameters returned more sequences than any of the others so far. Specifically we see any of the sequences returned by any of the other searches. We see a combination of the high similarity low E value sequences with the sequences who have lower identity and higher E values. It is also worth noting for the first two queries at least, the number of sequences with large E values is much large than the number of small E value sequences. This makes sense as there are only so many identical sequences, but when we relax the parameters we can pick up many more of the more distant sequences.

Summary

We can think about the two parameters we are investigating as impacting the results along similar axes. Decreasing the word size or using more general scoring matrices will drive E values up as the BLAST criteria are relaxed. However if we use for example the longest word size and most strict scoring matrix, we really only see identical sequences returned by the BLAST.